Development of Perception and Representation of Rhythmic Information: Towards a Computational Model

Inês Salselas, Perfecto Herrera
Universitat Pompeu Fabra
Music Technology Group
Barcelona, Spain

***Abstract*—We aim to model infants' perception and representation of temporal information that is present in infant directed speech and singing, using connectionist computational models (neural networks). In our approach, we consider the sound patterning, present in both speech and singing, in terms of timing and accent. The model receives audio on the input. Subsequently, different features are computed, according to different processes operating in parallel. Finally, we compute a representational transition, which learns categorical structured representations in terms of communication purposes from unstructured examples. In addition, we propose experiments to perform with the model. With these experiments we aim to study the development of representations from undifferentiated whole sounds to the relations between attributes that compose those sounds.**

Keywords-rhythm; music; speech; development; representation; computation.

I. EXTENDED ABSTRACT

Music and speech, from the perspective of a pre-verbal infant, may be perceived as sound sequences that unfold in time, following patterns of rhythm, stress and melodic contours. Therefore, it is likely that they share processing mechanisms and representation structures [1]. In this context, we aim to model infant's perception and representation of temporal information that is present in infant directed speech and singing. In our approach, we consider the sound patterning, present in both speech and singing, in terms of timing and accent. We leave the grouping or phrasal patterning for a later stage. The model receives audio on the input. Subsequently, different features are computed, according to different processes operating in parallel. Finally, we compute a representational transition, which learns categorical structured representations in terms of communication purposes from unstructured examples. The model has been conceptualized around two main modules: the *Perception Module* and the *Representation Module* (see Figure 1 and Figure 2).

In the *Perception Module* (see Figure 1), we hypothesize that for the perception of rhythmic information duration, intensity and pitch must be extracted from the audio input. Therefore, we have extracted this information from the vocalic intervals present in each input sound. Infants are able to segment vocalic intervals from the speech stream [2]. Vowels are perceptually relevant regarding rhythm since in languages with rhythmic patterns close to stressed-timing, which is the case, stress has a strong influence on vowel duration and the marking of certain syllables within a word as more prominent than others leads to vowels' duration fluctuation. In addition, the usage of vocalic intervals allows building a parallel with music since musical notes can be compare to syllables and vowels form the core of syllables [3].

We have used Prosogram [4] for segmenting vocalic intervals and extract duration, intensity and pitch values for each interval. This information that is extracted for each sound input, either speech or singing, will feed four elements that we called Contrast, Mode, Beat and Stress. Each of these elements will be described next.

- Contrast is based on the variability of the duration of the input units. For each input sound, nPVI is computed, which yields the index of contrast between neighboring durations. Contrast outputs "constant" for lower values of nPVI and "variant" for higher values of nPVI.
- Mode detects the input sound velocity by computing speech rate. Speech rate represents the number of durational units (vowels, in this case) per second. Mode outputs "fast" for higher speech rate values and "slow" for lower speech rate values.
- Stress identifies salient events in the input sounds. Salient events are marked whenever simultaneous peak values are spotted in Duration – Intensity or Pitch – Intensity pairs.
- Beat is fed by Stress and Contrast and detects regularity in the prominent events. Beat outputs "regular" for input sounds with beat and "irregular" for input sounds with no beat.

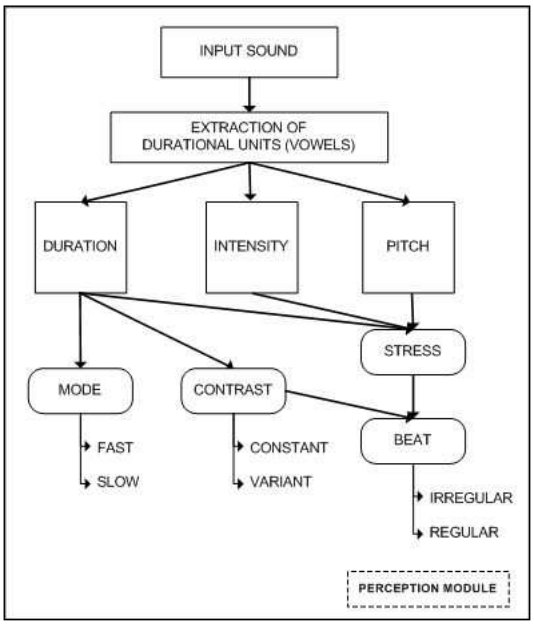

Figure 1– *Perception Module.*

Therefore, in our account, we employ a very primitive set of contrastive features which are fast/slow, constant/variant and regular/irregular. These features are extracted for each object that is presented as input sound in the *Perception Module.* All objects fit in to a high-level representation that is composed by four categorical objects: Prohibitive Speech, Affectionate Speech, Play-song Singing and Lullaby Singing (see Figure 2).

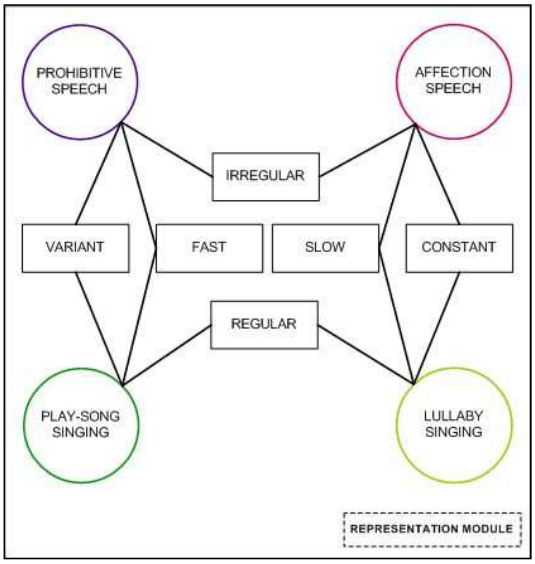

Figure 2 – *Representation Module*.

Similarly to the DORA model [5] that learns structured representations of relations from unstructured shape information related inputs, our model gradually discovers shared properties among different input sounds. Thus, the representation develops from a whole perspective of the input sound to a relational representation, building categorical relations between sound objects using the features extracted in *Perception Module* that compose the categorical objects.

Here, we present our initial experiments with this model. For these experiments, we take assumptions based on prior research: we assume that infants are able to detect regularity in prominent events in an auditory signal [6], that they can retain tempo [7] and that they can also discriminate languages based on the contrasts of their durations [8]. These assumptions were taken in order to define the basic properties (Slow/Fast, Constant/Variant and Irregular/Regular) assuring that the perceptual/cognitive system is able to detect them and thus, validate the *Perception Module* (see Figure 1). The sound materials that we use for the experiments consist of five elements from each prohibitive and affection speech and three elements from each play-song and lullaby singing. All of these elements have been taken from recordings capturing parents European Portuguese speakers, interacting spontaneously with their healthy babies aged up to 18 months.

With these experiments we aim to study the development of representations from undifferentiated whole sounds to the relations between attributes that compose those sounds. This transition should happen by means of consecutive comparison between features from each sound, discover shared features, build a relation between sounds through its common properties and assign to the sounds categorical meaning.

REFERENCES

[1] A. D. Patel, Music, Language, and the Brain, NY: Oxford Univ. Press, 2008.

[2] F. Ramus, M. Nespor and J. Mehler, "Correlates of linguistic rhythm in the speech signal", Cognition, vol. 73, no. 3, pp. 265-292, 1999.

[3] A. D. Patel, J. R. Iversen and J. C. Rosenberg, "Comparing the rhythm and melody of speech and music: The case of British English and French", Journal of Acoustic Society of America, vol. 119, no. 5, pp. 3034-3047, 2006.

[4] P. Mertens, "The prosogram: semi-automatic transcription of prosody based on a tonal perception model", In B. Bel & I. Marlien (eds.) Proceedings of Speech Prosody, Nara (Japan), 2004, pp. 549-552.

[5] L. A. A. Doumas, E. J. Hummel, "A computational account of the development of the generalization of shape information", Cognitive Science, vol. 34. No. 4, pp. 698-712, 2010.

[6] I. Winkler, G. P. Haden, O. Ladinig, I. Sziller, H. Honing, "Newborn infants detect the beat in music", Proceedings of the National Academy of Sciences of the United States of America, vol. 106, no. 7, pp. 2468-2471, 2009.

[7] L. J. Trainor, L. Wu and C. D. Tsang, "Long-term memory for music: infants remember tempo and timbre", Developmental Science, vol. 7, no. 3, pp. 289-296, 2004.

[8] T. Nazzi and F. Ramus, "Perception and acquisition of linguistic rhythm by infants", Speech Communication ,vol. 41, pp. 233-243, 2003.